Marginal Structural Models

applied to treatment for HIV infection

by

CHARLOTTE LIMA

THESIS

for the degree of

MASTER OF SCIENCE

(Modelling and data analysis)

Faculty of Mathematics and Natural Sciences

University of Oslo

April 2012

Det matematisk- naturvitenskapelige fakultet

Universitetet i Oslo

This thesis completes my Master's degree in Modelling and Data Analysis at the University of Oslo. Writing it would not have been possible without the endless support of my dear friends and precious family, not to mention the patience and encouragement from my supervisors Odd Aalen and Sven Ove Samuelsen. It has been a long and bumpy ride but I got there at the end and I am grateful for the experience and the knowledge I achieved on the way.

First I would like to thank my supervisors Odd Aalen and Sven Ove Samuelsen for the interesting discussions and great advice. Thank you for the patience and guidance.

The next goes out to my lovely family. My mother, her husband and my awesome little brother give me unconditional love and backup. The rest of the family "på Jæren" has also been there for me, thank you. Thanks to my godmother Miranda and her sister Adelaide for always believing in me and for correcting the "skriveleifs" in my thesis.

My friends have been my backbone all of the years studying at the University. Holding me up, in good times and bad. AnnaBanana, I have never met anyone like you. Your heart is filled with love, and you still have room for everybody in there. The rest of my study mates at Blindern Ella, Navreet, Joachim and the gang, we have had so much fun. Andréa, Jennie and the group of friends from primary and high school, still going strong. Plenty more dinners and TV shows to go. My friends in the salsa community and especially Eva and the coolest African girl I know Milgo. Girls, you make me laugh and forget about time when I am with you. Milgo, your open and warm family has a spirit that most of us lack. And to Jane. Your strength and courage is such an inspiration. Thank you.

Last but not least I want to thank my boyfriend Dan. Thank you for the endless tender, love and care.

I love you all.

Contents

CHAPTER 1

Introduction

Human Immunodeficiency Virus (HIV) is the virus that causes acquired immunodeficiency syndrome (AIDS). Being a member of a group of viruses called retroviruses, HIV infects human cells and uses the energy and nutrients provided by those cells to grow and reproduce. AIDS is a disease in which the body's immune system breaks down and is unable to fight off certain infections, known as "opportunistic infections", and other illnesses that take advantage of a weakened immune system. When a person is infected with HIV, the virus enters the body and lives and multiplies primarily in the white blood cells. These are the immune cells that normally protect us from disease. The hallmark of HIV infection is the progressive loss of a specific type of immune cell called T-helper or CD4 cells. As the virus grows, it damages or kills these and other cells, weakening the immune system and leaving the individual vulnerable to various opportunistic infections and other illnesses, ranging from pneumonia to cancer [1]. The U.S. Centers for Disease Control and Prevention (CDC), [2], defines someone as having a clinical diagnosis of AIDS if they have tested positive for HIV and meet one or both of these conditions:

- They have experienced one or more AIDS-related infections or illnesses
- The number of CD4 cells has reached or fallen below 200 cells per cubic milliliter (μL) of blood (a measurement known as T-cell count)

In healthy individuals, the CD4 count normally ranges from 450 to 1200 cells/μL.

For many years, there were no effective treatments for AIDS. Today, people in the United States and other developed countries can use a number of drugs to treat HIV infection and AIDS. Some of these are designed to treat the opportunistic infections and illnesses that affect people with HIV/AIDS. In addition, several types of drugs seek to prevent HIV from reproducing and destroying the body's immune system. Many HIV patients are taking several of these drugs in combination a regimen known as highly active antiretroviral therapy **(HAART)**. When successful, combination or "cocktail" therapy can reduce the level of HIV in the bloodstream to very low, even undetectable, levels and sometimes enable the body's CD4 immune cells to rebound to normal levels.

Researchers are working to develop new drugs known as fusion inhibitors and entry inhibitors that can prevent HIV from attaching to and infecting human immune cells. Efforts are also underway to identify new targets for anti-HIV medications and to discover ways of restoring the ability of damaged immune systems to defend against HIV and the many illnesses that affect HIV-infected individuals. Ultimately, advances in rebuilding the immune system in HIV patients will benefit people with a number of serious illnesses, including cancer, Alzheimer's disease, multiple sclerosis, and immune deficiencies associated with aging and premature birth [1].

With around 2.6 million people becoming infected with Human Immunodeficiency Virus in 2009, there are now (October 2011) an estimated 33 million people around the world who are living with HIV, including millions who have developed AIDS, [3]. Since the beginning of the epidemic, AIDS has killed nearly 19 million people worldwide. AIDS has replaced malaria and tuberculosis as the world's deadliest infectious disease among adults and is the fourth leading cause of death worldwide. In 2008 299 persons were reported diagnosed with HIV in Norway, and of these, 18 persons got AIDS. 12 of the persons with AIDS died because of it [4].

There is still no cure for AIDS, and while new drugs are helping many people with HIV/AIDS live longer, healthier lives, there are many problems associated with them [1]:

- Existing treatments do not work for many people with HIV/AIDS
- Anti-HIV drugs are highly toxic and can cause serious side effects, including heart damage, kidney failure, and osteoporosis. Many (perhaps even most) patients cannot tolerate long-term treatment with HAART
- HIV mutates constantly. In as many as 40% of people on HAART, HIV mutates into new viral strains that have become highly resistant to current drugs, and as many as 10% of newly infected Americans are acquiring drug-resistant strains of the virus
- Because treatment regimens are unpleasant and complex, many patients occasionally miss doses of their medication. Failure to take anti-HIV drugs on schedule and in the prescribed dosage can encourage the development of new viral strains that are resistant to current HIV drugs
- Even among those who do respond well to treatment, HAART does not eradicate HIV. The virus continues to replicate at low levels and often remains hidden in "reservoirs" in the body, such as the lymph nodes and brain

Importantly, roughly 95% of all people with HIV/AIDS live in the developing world, where there is virtually no access to antiretroviral treatments. In the U.S. HAART contributed to a significant decline in the annual number of AIDS-related deaths between 1996 and 1998. But the

rate of this decline has now slowed markedly, and some communities are reporting an increase in AIDS deaths.

When estimating the effect of a therapy such as HAART on progression to AIDS or death, proper methods for modelling is of great importance. Standard methods like Cox and logistic regressions in observational studies will fail to properly correct for the time-dependent confounders that are also affected by previous treatment. Thus the estimates from these methods will be biased. In the growing field of **Causality** there has been developed several methods to overcome this problem. One of them is the Marginal Structural Model **(MSM)**. The parameters of a MSM can be consistently estimated using the inverse-probability-of-treatment weighted **(IPTW)** estimators [5].

The aim of this thesis is the show the use of the Marginal Stuctural Model introduced by Robins [6]. The model will be presented in a biological context and applied to a training set of the Swiss HIV cohort [7] and some simulated data sets.

CHAPTER 2

Survival Analysis

2.1. Basic Concepts

Survival analysis deals with occurrences of events in different scientific studies, especially in medical research and statistics. Such events could be marriage, birth, graduation from school, failure in mechanical systems and so on. The term survival time doesn't necessarily have anything to do with death; it is the time from an initiating event to the time of the event of interest. The event of interest in my thesis is progress to AIDS or death and here the survival time is the time from treatment of HAART to the event.

Some subjects involved in a study will experience the event of interest, and others will maybe experience the event after the end of the study, or not even then. This leads to a censored data set consisting of complete and incomplete observations. The methods of survival analysis are specially developed for the handling of this kind of data.

2.1.1. Survival Function. The survival function specifies the unconditional probability that the event of interest has not happened by time t and is given by

$$S(t) = P(T > t)$$

where T denotes the survival time.

The survival function can also be looked at as the expected proportion of individuals for which the event has not happened by time t. Since more and more subjects experience the event over time the survival function decreases towards zero as t increases. In the situations where not all individuals experience the event, the random variable T may be infinite. Then $S(t)$ decreases to a positive number as T goes to infinity.

2.1.2. Hazard Rate. The hazard rate represents the instantaneous event rate for an individual who has already survived to time t and is given by

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t)$$

$\alpha(t)dt$ is the probability of experiencing the event in the small time interval $[t, t + dt)$.

Here it is assumed that T is absolutely continuous, meaning that it has a density function. $\alpha(t)$ can be any nonnegative function.

The relationship between $S(t)$ and $\alpha(t)$ is given by

$$\alpha(t) = -\frac{S'(t)}{S(t)}$$

The **hazard ratio** is the ratio of the hazard rates in two groups, for instance the group with treatment and the group without treatment. The hazard ratio is commonly used when presenting results in clinical trials involving survival data.

2.2. Cox Regression

The Cox model is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. It is the most commonly used multivariate approach for analyzing survival time data in medical research. The method is based on an assumption that the hazards remain proportionately constant and it is more correctly called the Cox proportional hazards model.

For the Cox model the hazard rate for individual i with vector of covariates $\mathbf{x_i(t)} = (\mathbf{x_{i1}(t)}, \ldots, \mathbf{x_{ip}(t)})^{\mathbf{T}}$ is given by

$$\alpha(t|\mathbf{x_i}) = \alpha_{\mathbf{0}}(\mathbf{t}) \exp\{\boldsymbol{\beta}^{\mathbf{T}} \mathbf{x_i(t)}\} \tag{2.1}$$

We have to look at a partial likelihood for the inference on the regression coefficients because of the semi parametric nature of (2.1).

Each individual i in the study has a counting process $N_i(t)$ counting the number of occurrences of the event of interest in $[0, t]$. $N_i(t) = 1$ if by time t the event has occurred for individual i, otherwise $N_i(t) = 0$. We assume that at time t for individual i the components of $\mathbf{x_i(t)}$ are fixed or time-varying, and that the intensity process of N_i can be written as

$$\lambda_i(t) = Y_i(t)\alpha(t|\boldsymbol{x}_i)$$

where $Y_i(t) = 1$ if individual i is at risk for the event just before time t and $Y_i(t) = 0$ otherwise. By using (2.1) the intensity process of N_i can be written as

$$\lambda_i(t) = Y_i(t)\alpha_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_i(t)\}$$

When registering events among all individuals we need the aggregated counting process $N_\bullet(t) = \sum_{l=1}^{n} N_l(t)$ with intensity process

$$\lambda_\bullet(t) = \sum_{l=1}^{n} Y_l(t)\alpha_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_l(t)\}$$

A factorization of the intensity process of $N_i(t)$ is given by $\lambda_i(t) = \lambda_\bullet(t)\pi(i|t)$ where

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda_\bullet(t)} = \frac{Y_i(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_i(t)\}}{\sum_{l=1}^{n} Y_l(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_l(t)\}} \tag{2.2}$$

is the conditional probability of observing an event for individual i at time t, given the past and that an event is observed at that time.

Assuming that we have no tied events, meaning that no events happen at the same time, we denote the times when events are observed by $T_1 < T_2 < \cdots$. The partial likelihood is obtained my multiplying the conditional probabilities (2.2) over all event times, and when i_j is the index of the individual experiencing the event T_j, it becomes

$$L(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j|T_j) = \prod_{T_j} \frac{Y_{i_j}(T_j)\exp\{\boldsymbol{\beta}^T \boldsymbol{x}_{i_j}(T_j)\}}{\sum_{l=1}^{n} Y_l(T_j)\exp\{\boldsymbol{\beta}^T \boldsymbol{x}_l(T_j)\}} \tag{2.3}$$

The $Y_{i_j}(T_j)$ in the numerator is always equal to one because the individuals experiencing the events are at risk just before the event times T_j. Defining the set of individuals who are still under study at time just before T_j as the risk set $\mathcal{R}_j = \{l|Y_l(T_j) = 1\}$,(2.3) can be written as

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{\exp\{\boldsymbol{\beta}^T \boldsymbol{x}_{i_j}(T_j)\}}{\sum_{l\in\mathcal{R}_j} \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_l(T_j)\}}$$

By maximizing the Cox log partial likelihood

$$l(\boldsymbol{\beta}) = \sum_{T_j} \left[\boldsymbol{\beta}^T \boldsymbol{x}_{i_j}(T_j) - \log \sum_{l\in\mathcal{R}_j} \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_l(T_j)\} \right]$$

we can estimate the regression coefficients in the Cox model. A positive coefficient will increase the hazard contributing negative on the survival, and a negative coefficient will decrease the hazard contributing positively in the survival.

The log hazard rate is given as

$$\log(\text{group hazard/baseline hazard}) = \log((h(t)/h_0(t)) = \sum_i \beta_i x_i \tag{2.4}$$

A unit increase in the independent variable i is associated with β_i increase in the log hazard rate.

2.3. Pooled logistic regression and the equivalence to Cox regression

In cohort studies in medical and epidemiological research a group of individuals is followed up for a study period of many years. Data about risk factors, an outcome of interest and other variables are collected repeatedly in set intervals of equal length over the study period. This repeated collection of data leads to several observations for each individual in the study. The interest is to evaluate the relationship between different risk factors to the outcome and the question is how to do this with the repeated observations.

The solution is to treat every interval as a mini follow-up study, pool the observations of all intervals together to one pooled sample, and do a logistic regression on the pooled data set. This is referred to as pooled logistic regression [8].

With this pooling method one single individual can contribute several times to the data set, or person-time. For each interval where the individual does not experience an event of interest or is not lost to follow-up for some reason, the individual can be carried on to be part of the risk set for the next period. It then counts as a new observation in the next interval.

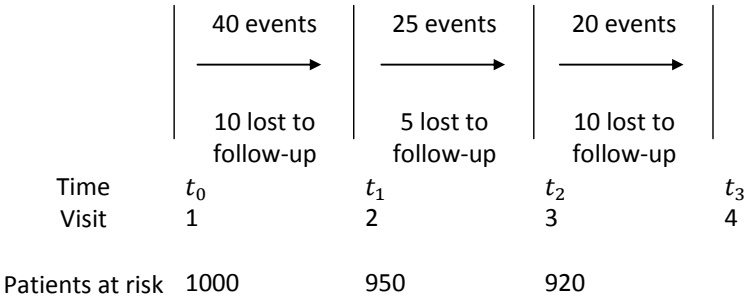

FIGURE 2.1. Pooling of repeated measurements

It is easier to explain the pooling method with the example that illustrates this type of pooled data from Figure 2.1. Imagine that 1000 patients enter a hypothetical study with 4 visits at times t_0, t_1, t_2 and t_3. Their risk factors are measured at the first visit and they are all at risk for the event at t_0. At the end of the first interval, 40 patients may have experienced the event, and 10 may have been lost to follow-up. The 40 events are recorded, and the 50 patients experiencing the event or being lost to follow-up are removed from the risk set. The remaining 950 patients are measured at the second visit carried on forming the risk set at time t_1. At the end of the interval starting at time t_1, there has been 25 events and 5 patients lost to follow-up. These 30 patients are removed from the risk set, leaving 920 patients at risk at time t_2 with risk factors being measured at the third visit. From t_2 to t_3 there has been 20 events and 10 patients lost to follow up. So the 1000 patients who entered the study contributed with 1000+950+920 = 2870 person-time as if they were individual observations. In total there were 40+25+20=85 events. The pooled data set consists of 2870 observations with 85 events, and a logistic regression is performed on this pooled sample.

With very short intervals, the probability of an event happening in an interval is very small and the intercept for the pooled logistic regression is constant. When this is the case the model of pooled logistic regression is asymptotically equivalent with the Cox time-dependent regression model [9], meaning a Cox regression model with time-dependent covariates.

The logistic regression model is written:

$$\text{logit}q_i(\mathbf{X}(t_{i-1})) = \log\left(\frac{q_i(\mathbf{X}(t_{i-1}))}{1-q_i(\mathbf{X}(t_{i-1}))}\right) = \alpha_i + \beta_1 X_i(t_{i-1}) + \ldots + \beta_p X_p(t_{i-1}),$$

where $q_i(\mathbf{X}(t_{i-1}))$ is the conditional probability of observing an event by time t_1 given that the individual is event free at time t_{i-1}, and $\mathbf{X}(t_{i-1}) = (X_1(t_{i_1}), \ldots, X_p(t_{i_1}))$ is the risk factors measured at time t_{i-1}. The intercept α_i is a function of the time between visit $i-1$ and visit i.

When the risk factors are recorded at times $t_0, t_1, \ldots$ (Figure 2.1) and the observations are grouped into intervals $[t_{i-1}, t_i]$ then the hazard rate in the time dependent covariate cox regression is given:

$$p_i(\mathbf{X}(t_{i-1})) = \exp\left\{-\int_{t_{i-1}}^{t_i} h_0(u)\exp[\beta'\mathbf{X}(t_{i-1})]du\right\}.$$

$p_i(\mathbf{X}(t_{i-1}))$ denotes the probability that an individual will survive up to time t_i given survival up tp time t_{i-1}. $h_0(u)$ is the baseline hazard rate and $\beta'\mathbf{X}(\mathbf{t_{i-1}}) = \beta_1 X_1(t_{i_1}) + \ldots + \beta_p X_p(t_{i_1})$ is the linear function of the Cox proportional hazard model. In this model the events are grouped into intervals $[t_{i-1}, t_i]$ but not specified as to a time of occurrence within the intervals. This is the grouped Cox model.

CHAPTER 3

Causality

Causality is based on the notion of the past influencing the present and the future [10]. Underlying the concept of causality is the simple word *cause* and in questions of causal matters the relationship between cause and effect is explored closely. When practising causal inference, it is important to bear in mind that there is a difference between association and causation. Just because studies in children have revealed that shoe size is positively *associated* with literacy, it does not automatically imply that becoming more literate makes your feet grow or vice versa. The groups of children that are compared may not be comparable because they differ in terms of factors other than their literacy score, age for instance.

In life we are exposed to interventions for the achievement of different goals. It could be a medicine prescribed from a doctor to cure a disease. The medicine is supposed to have a specific causal effect on the disease. Before the medicine is put on the marked, it is put through a long process of research and clinical testing. We need some understanding of how for example the body or viruses respond to certain substances. What lies behind causality is the search for a mechanistic understanding of connections surrounding specific factors. This mechanistic understanding is very limited when it comes to medicine because of the complexity and the countless number of causal connections in the body. In some cases there are some ethical limitations to the interventions and experiments, but the aim to understand the effect of the interventions is still there. In other cases we know the causal pattern already because some variables are known earlier than others. University grades will for instance have no impact on high school grades, but rather the other way around.

We want a clear overview of the causal connections between the factors involved in a study not to make the mistake of misinterpreting associations and direct effects. The connections between the variables can be summed up using graphical models that show how the different variables influence one another. One example of a graphical model is the DAG, **directed acyclic graph**.

3.1. DAG

A causal diagram is a system of nodes and edges, based on expert knowledge that represents all causal influences between all relevant observed variables. With this graphical representation we can gain insight whether the effect of one variable on another is identified and how it can be

estimated. An example of a causal diagram is a DAG, Directed Acyclic Graph. The DAG is a system of directed edges (arrows) between variables but without cycles. Each edge on the DAG represents a *possible* direct causal relationship between two variables, and the absence of an edge represents the assumption that there is no direct causal relationship between the variables. All common causes of any two variables are included as a variable on the DAG. Thus a causal DAG does not have to include variables that are not of interest and not a common cause of two variables in the DAG. In other words, there are no omitted confounders in the DAG.

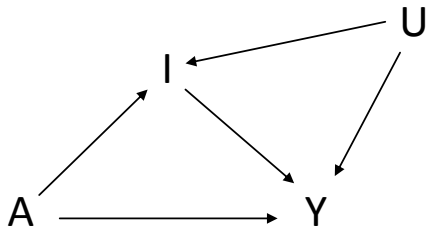

FIGURE 3.1. DAG

Figure 3.1 is an example of a DAG with only one time point. A can be an exposure of some kind and Y a specific outcome. Then I could be a vector of all measured risk factors for Y, and U all unmeasured risk factors for Y. One variable can only causally affect another variable when there is a directed path between the two. In the example of Figure 3.1 A may have a direct causal effect on Y as well as an indirect causal effect which is mediated by I. A does not causally affect Y along the path $A - I - U - Y$. But two variables can be associated along all directed and undirected paths that connect the two. With a graphical rule called d-separation, we can decide whether a conditional association between measurements reflects causation under the assumptions of the causal diagram.

3.1.1. d-separation. d-separation is a graphical rule to verify independence between variables based on a DAG. A **collider** is a node with two or more arrows pointing to it. If we think of the DAG as an electric circuit, the colliders are switches that are turned off (inactive), and the non-colliders are switches that are turned on (active). If there is no electric current between two variables they are independent. There may be association between two variables along all active paths. The association between A and Y in Figure 3.1 is due to the direct causal effect and the indirect causal effect through I. But the path $A - I - U - Y$ is inactive because of the collider I and does not associate A and Y. If we do an analysis and restrict the analysis to subjects with the same value of I or include I in a regression model $E(Y|A, I) = \alpha + \beta A + \gamma I$, we "adjust the analysis for I". With the rules of d-seperation we can evaluate whether A and Y

are conditionally independent. Adjusting for non-colliders changes them from active to inactive and adjusting for colliders or their descendants changes them from inactive to active. If there is no electric current between A and Y after adjusting for I, they are conditionally independent given I.

For example adjusting for I in Figure 3.1 closes the $A - I - Y$ path but opens the $A - U - Y$ path. The conditional *association* between A and Y is thus due to the direct causal effect and the spurious association through the unobserved U, but not due to the indirect causal effect through I. It's important not to ignore the possible presence of confounders for the association between mediator and outcome. The approaches of a traditional mediation analysis with the adjustments are invalid in the presence of confounders U. The reason for this is that they try to uncover causation merely from statistical associations, but **association $\neq$ causation**. Thus it is very important to express background knowledge when we want to learn about the effect of some exposure on some outcome. Using d-separation, we can infer for which confounders we need to adjust when estimating the causal effect of A on Y.

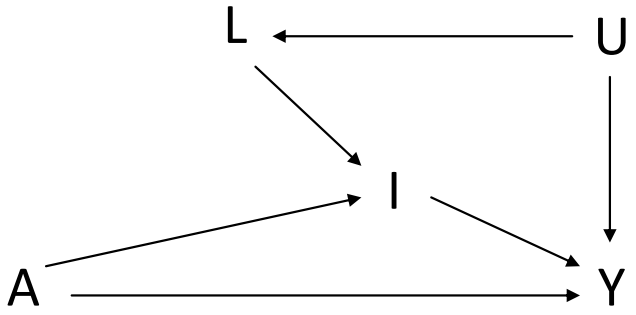

FIGURE 3.2. DAG

Conditional association between A and Y in Figure 3.2 given I and L reveals the direct effect of A on Y. Thus when we fit model

$$E(Y|A, I, L) = \alpha + \beta A + \gamma Y + \delta L$$

then β measures the **direct** effect of A on Y.

3.2. Time-dependent confounding and complications

A confounder is a common predictor of two or more variables and it is time-dependent when it varies with time. In this case with the observational study of HIV patients, looking on the effect

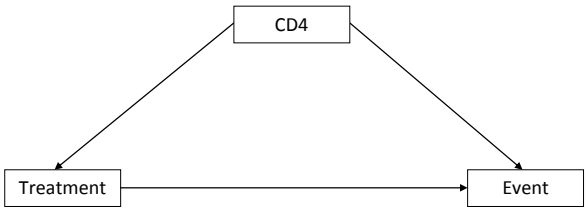

FIGURE 3.3. Confounding between treatment and event

of a treatment (exposure) on time to AIDS or death (event), the time-dependent confounder has an impact on both the event and also on treatment. One example of this type of confounder is the CD4 count. The CD4 cells are one of the different types of cells that help protect the body from infection, and the CD4 count, among other factors, implies the severity of the HIV disease. The higher the number of CD4 cells in the blood the better the immune system. HIV attacks these types of cells and weakens the immune system. An HIV patient is classified as having AIDS when the CD4 count drops below a certain limit. So the CD4 count affects the event. The CD4 count is also used in the decision of when to initiate the HIV treatment. Thus the CD4 count confounds the relationship between treatment and the event, see Figure 3.3 for this simple example with only one time point.

The reality is that CD4 count is also affected by treatment and this is when we get a feedback relationship between treatment and CD4, they both affect each other. The CD4 is a confounder of the relationship between treatment and event, but is also a mediator on the causal path from treatment to event. Thus CD4 is an **intermediate time-dependent confounder** and with this follows complications.

A bit more complex and general example of a DAG is given in Figure 3.4. Here possible relationships between the variables are given with two time points, $k = 1, 2$.

- A_k is the exposure at time k
- Y is the outcome
- L_k is a vector of all measured risk factors for the outcome at time k
- U_k is all unmeasured risk factors for the outcome at time k

From Figure 3.4 we can see that L is an intermediate time-dependent confounder of the relationship between exposure and outcome. The causal effect of exposure on outcome is divided on four different directed paths: $A_0 - A_1 - Y$, $A_1 - Y$, $A_0 - L_1 - A_1 - Y$ and $A_0 - L_1 - Y$, describing direct and indirect effects of exposure. To model this effect we could fit a standard regression model adjusted for exposure A:

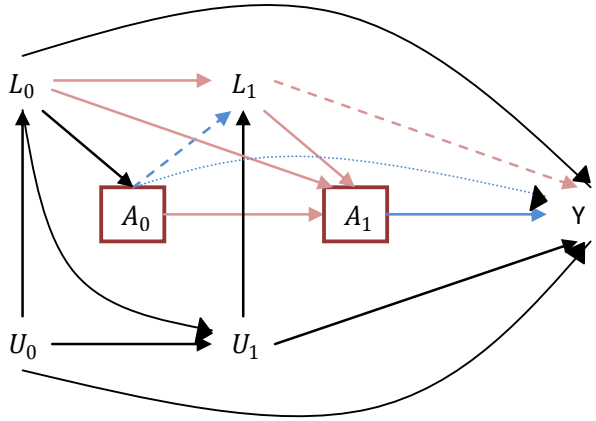

FIGURE 3.4. Intermediate time-dependent confounding

$$E(Y|A_1, A_0) = \omega + \beta_0 A_0 + \beta_1 A_1$$

The setting of Figure 3.4 together with the rules of d-separation, explained in subsection 3.1.1, shows that adjusting for A_0 and A_1 opens the undirected paths through L_0 and L_1. Thus the estimated exposure effect from the standard regression model is not the causal effect that we wish to estimate. We could try to adjust for L_0:

$$E(Y|A_1, A_0, L_0) = \omega(L_0) + \beta_0 A_0 + \beta_1 A_1$$

Even though this closes one of the non-causal paths through L_0 there is still the undirected path from exposure going through L_1. So the problem from the first standard model is the same. We could adjust for L_1 in the regression model as well:

$$E(Y|A_1, A_0, L_0, L_1) = \omega(L_0, L_1) + \beta_0 A_0 + \beta_1 A_1$$

Again with the rules of d-separation we see that this not only closes parts of the causal effect, going from A_0 along L_1, but also opens the spurious association between A_0 and Y through U_1. As this variable is unmeasured, it cannot be adjusted for.

The problem with these standard approaches is that they all fail to reflect the causal effect of exposure. We need to adjust for L as it confounds the $A-Y$ relationship, but we are not allowed to do this adjustment because it removes part of the causal effect and also creates a spurious association. The problem of the possible presence of unmeasured factors U, that are associated with outcome and the confounders, needs to be taken into account.

3.3. Marginal Structural Models

Marginal Structural Models (MSM) belong to a class of causal models used for the estimation of the causal effect of a time-dependent exposure in the presence of time-dependent covariates that are themselves affected by previous treatment. The effect is modelled using counterfactual outcomes. The MSM is *marginal* because it describes the effect of the exposure on the marginal distribution of the counterfactual outcomes and *structural* because models for counterfactual random variables are called structural in the social and economic sciences [11]. The parameters of the MSM can be estimated using inverse-probability-of-treatment weights. Valid causal inferences can be drawn by comparing the subjects who are on treatment with the patients who are off treatment in different time intervals. Thus the same patients represent observations in different treatment groups.

The model is fitted in two stages [12]:

(1) estimation of each subject's probability of having their own treatment history and calculation of inverse-probability-of-treatment weights (IPTW)
(2) the effect of treatment is estimated in a regression model that is weighted using the IPTW's

The marginal structural model is based on some assumptions:

(1) there are no unmeasured confounders
(2) the marginal structural model for the effect of HAART on AIDS or death among the HIV patientes is correctly specified
(3) the model for initiation of treatment is correctly specified

3.3.1. Counterfactuals. The concept behind counterfactual thinking is to imagine what might happen if possibilities other than the actual one did occur. We look at an *underweight* new born child whose mother smoked during the pregnancy. What would have happened with the weight of the child hadn't the mother smoked? This possibility is "counter to fact". In this case we can only observe one of the outcomes of the "treatment" (smoking), but the idea is to mimic the other outcome and compare both possibilities.

We denote $T_{\bar{a}}$ as the subject's time to event (e.g failure time) had he received treatment history $\bar{a}$ rather than his observed history. We only observe the failure times where $T_{\bar{a}} = T$ where T is the observed time to event, the other values of $T_{\bar{a}}$ being counterfactuals. For each $\bar{a}$ we specify the marginal structural Cox proportional hazards model

$$\alpha_{T_{\bar{a}}}(t|V) = \alpha_0(t)\exp\{\beta_0 a(t) + \beta_1 V\}$$

where $\alpha_{T_{\bar{a}}}(t|V)$ is the hazard of the event at time t among subjects with baseline covariates V had they all followed treatment history $\bar{a}$. β_0 and β_1 are the unknown parameters to be estimated and α_0 is an unspecified baseline hazard function.

3.3.2. Inverse-probability-of-treatment (IPT) weights. A pseudo population in which the risk factors no longer confound the relationship between treatment and outcome is created using IPT weights. Each subject gets weighted according to the subject's probability of having its observed treatment given the observed covariates for every time point. This way the subjects contribute to the new risk set with copies of themselves. The rare cases get up-weighted and the most common cases get down-weighted. The initiation of treatment in the reweighted set is no longer dependent on the risk factors. An example of a rare case is a subject with small probability of getting treatment who starts treatment. An example of a common case is a subject with big probability of starting treatment who starts treatment.

It is common to define the weights based on a discretization of the time interval. The IPT weight for subject i is given by

$$w_i(t) = \prod^{int(t)} \frac{1}{P(A(k) = a_i(k)|\bar{A}(k-1) = \bar{a}_i(k-1), \bar{L}(k-1) = \bar{l}_i(k-1))} \tag{3.1}$$

The probability of starting treatment may vary greatly among the subjects leading to unstable weights. A few subjects may get very large weights and with this contribute to the pseudo population with a large number of copies of themselves. These subjects will dominate the weighed analysis.

The stabilized version of inverse-probability-of-treatment weight is given by

$$sw_i(t) = \prod^{int(t)} \frac{P(A(k) = a_i(k)|\bar{A}(k-1) = \bar{a}_i(k-1), V = v)}{P(A(k) = a_i(k)|\bar{A}(k-1) = \bar{a}_i(k-1), \bar{L}(k-1) = \bar{l}_i(k-1))} \tag{3.2}$$

where $\bar{A}(-1) = 0$. This results in weights that are less variable, are centered around 1 and are closer to the normal distribution.

Each factor of the denominator of $sw_i(t)$ is the probability that the subject received his own observed treatment at month k, given his past treatment and prognostic factor history, where V is included in $L(0)$. Each factor in the numerator is the probability that the subject received his own treatment conditional on his past treatment history and baseline covariates, but not for the time-dependent confounders.

Weighting by $sw_i(t)$ creates the pseudo population we want where $\bar{L}(t)$ no longer predicts the initiation of HAART at time t, that is, $\bar{L}(t)$ is no longer a confounder of the treatment-event relationship.

CHAPTER 4

A training set provided by Jonathan A. C. Sterne

The Swiss HIV Cohort Study is an on-going multi-center research project in Switzerland that was established in 1988, dealing with HIV infected adults aged 16 years and older [7]. After HAART was introduced in Switzerland in 1996, Sterne et al [13], published an article with a study looking at the effectiveness of antiretroviral therapy in preventing AIDS and death using data from the Swiss HIV Cohort Study. The estimation of the treatment effect is calculated using the method of MSM. Jonathan A. C. Sterne has been so kind to provide a training set consisting of a small part of the data set used in the study. With this training set comes a sheet of commands of how to analyze the data set in the statistical software Stata . Since I am not very familiar with this program I use this training set and methods in Stata to create an equivalent program in R. I compare the results from the two programs to make sure that I have the right programming base in R for further demonstration of the MSM.

The event of interest in the training set is AIDS or death. The training data set is split into monthly intervals and contains the following information about the patients for each month:

- Cd4 - current CD4 group - lowest CD4 (≤ 50) is the reference
- Rna - current plasma HIV-1 RNA group - highest RNA ($\geq$ 100,000) is the reference
- LCd4 - lagged (three months previous) CD4 group
- LRna - lagged (three months previous) plasma HIV-1 RNA group
- BCd4 - baseline CD4 group
- BRna - baseline (three months previous) plasma HIV-1 RNA group
- A - age group
- Y - current year (from 1996, with 2001 as the reference year)
- Group - transmission risk group (with men who have sex with men (MSM) as the reference group)

There are a lot of similar lines for each patient. If there is no new observation for one month, the observation from the previous month is used.

Patients were exluded from the study if they:

- died or refused further participation before 1996
- were on HAART or had the AIDS diagnosis at the first follow up visit

- had an uncertain treatment history before joining the study

The data set contains measurements of CD4 count, HIV-1 RNA, hemoglobin, information of which treatment was taken (monotherapy, dual therapy or HAART, and information on whether the patients experienced a CDC stage B, i.e. got a CD4 count from 200-499 cells/μL [2]. The CD4 count is as desbribed in Chapter 1 a count of a specific type of immune cells in the body. The HIV-1 RNA is the RNA copies per millimeter of blood plasma. The CD4 count and HIV-1 RNA are used as markers for the severity of the HIV infection [15]. Hemoglobin in the blood carries oxygen from the lungs to the rest of the body where it releases the oxygen to burn nutrients to provide energy to power the functions of the organism, and collects the resultant carbon dioxide to bring back to the lungs to be dispensed from the organism [16].

The first monthly visit after January 1996 for which each variable was available was the baseline month. To get more conservative estimates it was assumed that treatment was started at the end of the month before they actually started the treatment. Another assumption was that once treatment was started the patient stayed on it.

I do unweighted Cox analyses in Stata with the recipe I received from Sterne using:

(1) treatment as the only covariate
(2) treatment and baseline covariates
(3) treatment, baseline, and time-updated covariates

Analysis	**HR**	**SE**	**P-value**	95% **C.I**
(1)	0.75	0.20	0.14	$[0.51, 1.10]$
(2)	0.37	0.22	0.00	$[0.24, 0.56]$
(3)	0.72	0.23	0.15	$[0.45, 1.13]$

TABLE 4.1. Unweighted Cox analyses in Stata

The only significant result from Table 4.1 is the analysis where treatment and baseline covariates are used. It shows that the hazard for AIDS or death is reduced with 63% for the patients on treatment compared to the patients not on treatment. It is clear that HAART has an effect.

Analysis	**HR**	**SE**	**P-value**	95% **C.I**
MSM	0.15	0.22(0.08$*$)	0.00	$[0.12, 0.23]*$

TABLE 4.2. Weighted Cox analysis in Stata

*Estimated from boostrap sampling.

The overall hazard ratio from the weighted Cox analysis in Table 4.2 is $0.15(0.12, 0.23)$ for treatment compared with no treatment. This result is significant and shows a much stronger effect of treatment than the unweighted analyses. The hazard for AIDS or death is now reduced with 85% for the patients on treatment. This result coincides with the overall hazard ratio in the study by Sterne et al [13], even though the data set used here is much smaller than the data set in the study.

The MSM gives a more realistic estimate of the treatment effect than the standard unweighted analyses, especially compared to the significant result of the unweighted Cox where treatment and baseline covariates were used. The fact that the patitients starting treatment generally have a severe HIV infection, and therefore a bigger chance of getting AIDS or dying, is not considered in the unweighted analyses. The intermediate confounding is controlled for using the MSM.

4.1. Effect of lagged variables

Analysis	**HR**	**SE**	**P-value**	95% **C.I**
MSM2	0.14	0.22	0.00	[,]

Table 4.3. MSM without lagged covariates of CD4 and RNA

The lagged variables of CD4 and RNA, i.e the measurements three months before the current month, are used in the weighted analysis. It is interesting to check if the effect of treatment changes when the lagged varibles are not used as covariates in the analysis. This is done in Table 4.3. The hazard ratio is 0.14 compared to 0.15 as is the case when the lagged variables are used. This is certainly not a big difference, so the lagged variables seem to have a tiny impact on the effect estimate of HAART.

4.2. Stata/R

The making of the equivalent program in R was not too hard to do because of the analysis tools in R. But the handling of two or more events in the same time interval, e.g ties, were different in the two programs. The default method in Stata is the method of Breslow and the default method in R is Efron. I chose Breslow's method for the handling of ties. The event covariate "aidsordeath" in the data set is either 0, 1 or ".". The dot means a censored event. The standard procedure in R for missing data in a Cox analysis is to delete the lines with missing data, or a ".". To avoid the deleted lines in R, I put all the dots in the event covariate "aidsordeath" to zero and made an new indicator variable for censoring.

Analysis	**HR**	**SE**	**P-value**	95% **C.I**
(1)	0.75	0.20	0.13	$[0.51, 1.10]$
(2)	0.36	0.22	≤ 0.001	$[0.24, 0.55]$
(3)	0.71	0.23	0.14	$[0.45, 1.12]$

TABLE 4.4. Unweighted Cox analyses in R

Analysis	**HR**	**SE**	**P-value**	95% **C.I**
MSM	0.17	0.21	≤ 0.001	$[0.11, 0.26]$

TABLE 4.5. Weighted Cox analysis in R

The results in Table 4.4 and Table 4.5 show small differences from the results in Stata, Table 4.1 and Table 4.2. The reason for the small differences is that Sterne has used splines to model the change in the hazard with time. I have not done this in R. Except from the use of splines I have made an equivalent program for the analyses in R and this program is used for further analyses.

Table 4.6 gives a summary of the most important predictors for starting HAART.

	HR	**P-value**
Transmission group		
Group 1	1 (reference)	
Group 2	1.14	< 0.001
Group 3	0.57	< 0.001
Group 4	1.73	< 0.001
CD4 count		
< 50	1 (reference)	
$50 - 99$	0.89	0.54
$100 - 199$	0.63	0.01
$200 - 349$	0.41	< 0.001
$350 - 499$	0.36	< 0.001
$500 - 749$	0.41	< 0.001
≥ 750	0.46	< 0.001
Lagged CD4 count		
< 50	1 (reference)	
$50 - 99$	0.92	0.70
$100 - 199$	0.73	0.13
$200 - 349$	0.75	0.19
$350 - 499$	0.64	0.05
$500 - 74$	0.65	0.05
≥ 750	0.91	0.67
RNA copies		
< 400	35.72	< 0.001
$400 - 1000$	8.17	< 0.001
$1'001 - 10'000$	2.48	< 0.001
$10'001 - 100'000$	1.30	< 0.001
$> 100'000$	1 (reference)	
Lagged RNA copies		
< 400	3.28	< 0.001
$400 - 1000$	1.27	0.01
$1001 - 10'000$	0.88	0.08
$10001 - 100'000$	1.01	0.85
$> 100'000$	1 (reference)	

TABLE 4.6. Effect of covariates on starting HAART

CHAPTER 5

Simulation and analyses

To show the use of the Marginal Structural Model I simulate a set of patients already diagnosed with HIV followed for a period of three years.

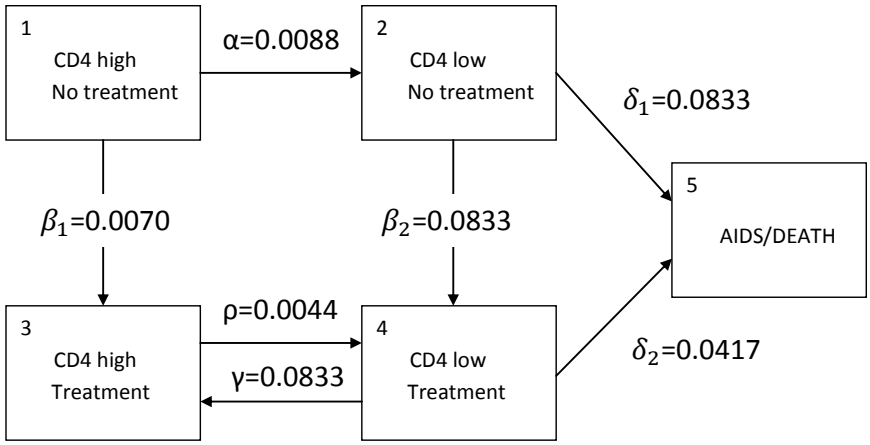

FIGURE 5.1. Markov model

This is a very simple model with CD4 and treatment as the only covariates. The covariates are used as information to describe the course of the disease, and the different stages follows a multi-state Markov model with the state space and intensities given in Figure 5.1. The Greek letters are the transition intensities between the different states of the model. The patients can start in state 1, 2, 3 or 4. The model is based on the model for the development of AIDS and HIV diagnosis of Aalen et al from 1997, [17], but very simplified. Towards the right of Figure 5.1 is the progression of the disease, ending with either AIDS or death being the same state. Downwards is the start of a drug treatment that can be started in both stages of the disease. The model is homogeneous over time to make it as simple as possible, meaning that the transitions don't depend on the times explicitly but on the length of the time interval. It is assumed that if a patient starts treatment he stays on it. And it is not possible to move from low CD4 to high CD4 without treatment.

The two stages of the HIV progression are

- Stage I: CD4 $\geq$ 200 cells/μL, states 1 and 3
- Stage II: CD4 $<$ 200 cells/μL, states 2 and 4

The occupancy is the percentage of time spent in a state against the (AVAILABLE TIME??). The mean occupancy time, T, is then the mean time spent in the state. In the thesis of Odd Aalen [17] there were used three stages with mean occupancy times 5.5 years, 4 years and 1 year. My idea is to sort of merge two of those stages together, resulting in two stages with mean occupancy times 9.5 years and 1 year. The data set is split into monthly intervals and then the transition intensity β is given as

$$\beta = \frac{1}{12 \times T}$$

This gives the transition intensities $\beta_1 \approx 0.0070$ and $\beta_2 \approx 0.0833$.

The other parameters are fixed so that:

- the intensity of starting treatment is almost 12 times bigger with a low CD4 count than with a high value
- the intensity for getting a low value of the CD4 count is two times bigger when off treatment than on
- the intensity for getting AIDS or dying is two times bigger when off treatment than on

γ was set to a value that seemed reasonable.

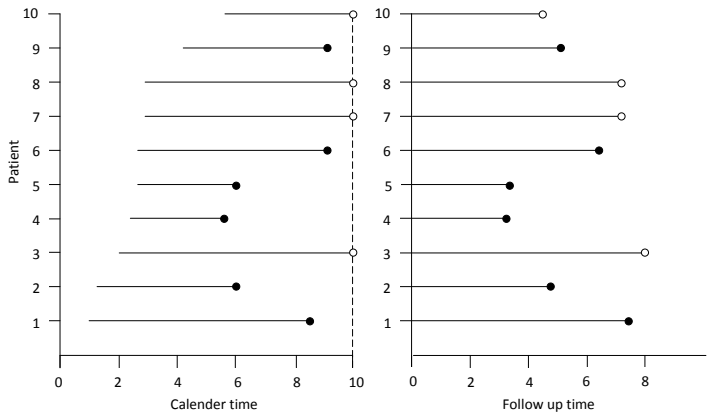

FIGURE 5.2. Follow up

When individuals enter the study at different times, it causes left-truncation and when the study is stopped at a certain time we get censored event times. Left-truncation is basically a delayed entry, and censoring is missing data. We can summarize these calender time data to follow up time. Figure 5.2 is a hypothetical clinical study with 10 patients. The filled circles indicate the occurrence of the event, and the open circles indicate censoring. The follow uptime is used in my simulation.

$$Q = \begin{vmatrix} -0.0158 & 0.0088 & 0.0070 & 0 & 0 \\ 0 & -0.1667 & 0 & 0.0833 & 0.0833 \\ 0 & 0 & -0.0044 & 0.0044 & 0 \\ 0 & 0 & 0.0833 & -0.1250 & 0.0417 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

FIGURE 5.3. Intensity matrix

The 5×5 intensity matrix Q for the course of the patients is given in Figure 5.3. The rows of Q must sum up to zero, so the diagonal elements are the negative sum of each row.

From Markov theory [18] it is possible to find an explicit solution of the transition matrix $P(t)$ for a finite Markov chain given the intensity matrix Q. The theory is that $P(t) = e^{Qt}$ where e^{Qt} is called the **matrix exponential**.

The transition probabilities are defined as the probability of occupying state j at time t conditional on occupying state i at time s.

$$p_{i,j}(s,t) = Prob\{X(t) = j | X(s) = i\}, \quad s < t$$

for $i, j = 1, \ldots, 5$. But we have a time-homogeneous chain, and the probabilities are then denoted as $p_{i,j}(t)$, being the probability of going from state i to state j in time t. These probabilities are given as the (i, j) entries of $P(t)$. For instance we have

$$P(1) = \begin{vmatrix} 0.9843 & 0.0080 & 0.0070 & 0.0003 & 0.0003 \\ 0 & 0.8465 & 0.0031 & 0.0720 & 0.0783 \\ 0 & 0 & 0.9958 & 0.0041 & \approx 0 \\ 0 & 0 & 0.0782 & 0.8827 & 0.0392 \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix}$$

FIGURE 5.4. Intensity matrix

These are the probabilities of moving between the states in time 1.

The survival matrix at time t is

$$S(t) = 1 - P(t)$$

5.1. Theoretical analyses

I use R to find the transition matrices for 100 time points between 1 and 36 months and compare the situation where the patients get a treatment offer and the situation where there is no treatment offer. The first situation is given with the model in Figure 5.1 and the second with $\beta_1 = \beta_2 = 0$.

Figure 5.8 gives the probabilities of going from state 1 to state 5 for the setting of a treatment offer and for the setting of no treatment offer. The probability of going from state 1 to state 5 starts out the same for the two different settings until about month 5. From then the probability is higher when there is no treatment offer than when there is one. This difference gets bigger and bigger as time goes. Now I have a base for comparison of the results I get from further analyses.

Figure 5.9 gives the corresponding survival function. It starts at 1 and then decreases slowly.

Figure 5.10 shows a decreasing hazard ratio, meaning that there is a treatment effect and that it gets better and better with time.

5.2. Simulation of patients with random start

In the simulated data set there is 244 events and 756 censored events for the period of 3 years. The set is simulated in continuous time.

(1) treatment as the only covariate
(2) treatment and baseline covariates
(3) treatment and time-updated covariates

5.2.1. Unweighted Cox Analysis. Table 5.1 shows the results for the Cox Analyses for the random start set. There is a significant positive treatment effect for all three of the analyses. For analysis no. 1 there is a decreased hazard for AIDS or death of 66% with the treatment compared to no treatment. For analalysis no. 2 the hazard is reduced with 88% for the patients on treatment. The hazard for the patients on treatment in the analysis no. 3 is 74% lower than

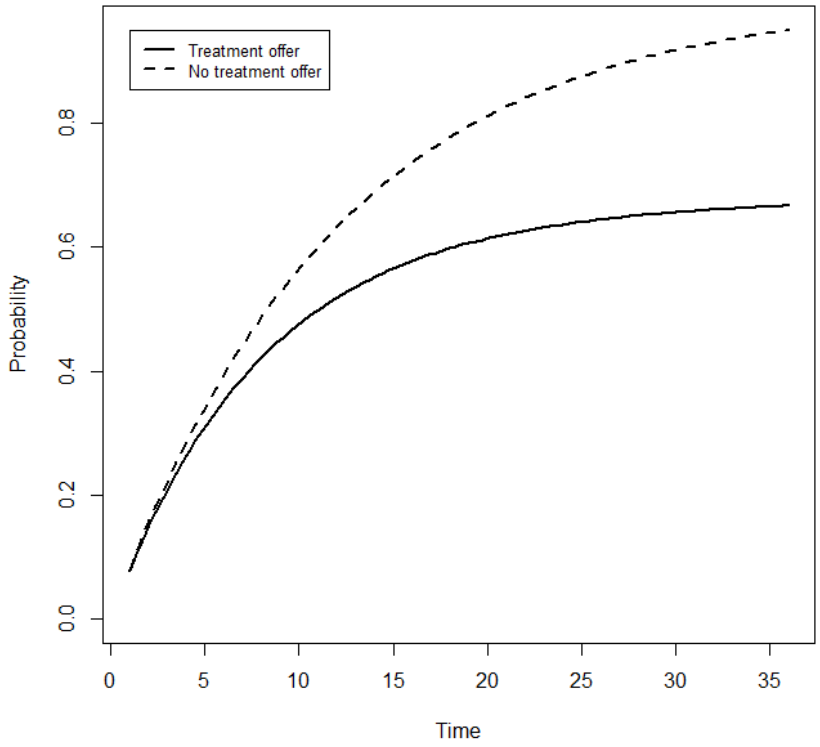

FIGURE 5.5. $p_{2,5}(t)$: Probabilities of moving from state 2 to state 5, with and without a treatment offer

Analysis no.	**HR**	**SE**	**z**	**P-value**	95% **C.I**
(1)	0.3406	0.1247	-8.636	≤ 0.001	$[0.2667, 0.4349]$
(2)	0.1229	0.1859	-11.278	≤ 0.001	$[0.0854, 0.1770]$
(3)	0.2618	0.1767	-7.582	≤ 0.001	$[0.1852, 0.3702]$

TABLE 5.1. Unweighted Cox Analysis

for the no treatment group. The last hazard ratio is not correct because of the warning: "Loglik converged before variable 3; beta may be infinite". This means that there is too little variation in variable 3 which is the cd4 variable.

Analysis	**HR**	**SE**	**z**	**P-value**	95% **C.I**
Weighted	0.2571	0.1263	-10.7553	≤ 0.001	$[0.2008, 0.3293]$

TABLE 5.2. Results from the weighted analysis, random start state

5.2.2. Weighted Cox Analysis. Table 5.2 shows that the odds for AIDS or death is reduced with 74% for the patients on treatment compared to those off treatment. This is a very good treatment estimate. The hazard ratio is a little higher than the one calculated by Sterne

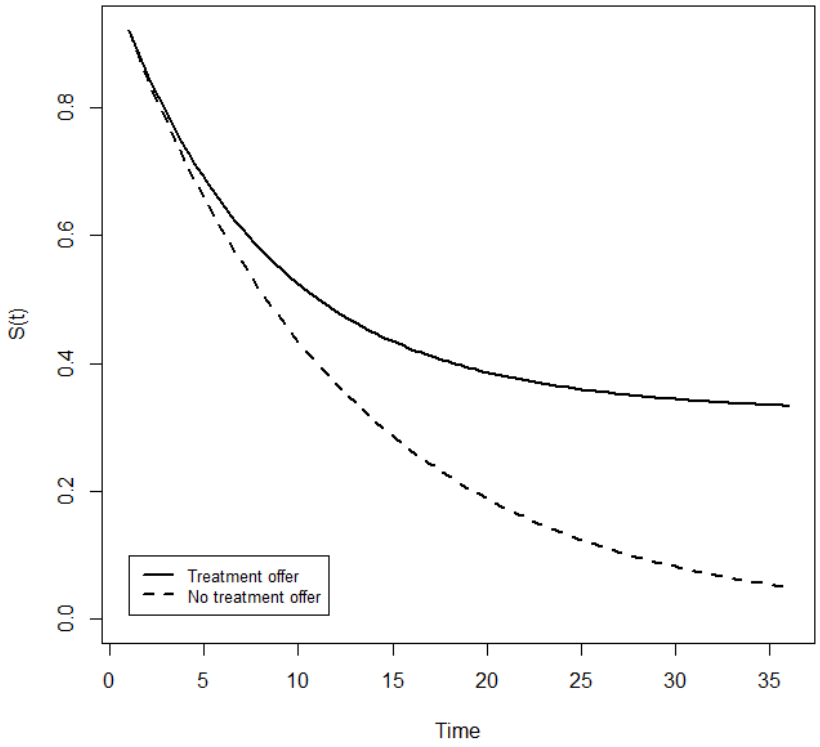

FIGURE 5.6. $S_{2,5}(t)$: Probabilities of not moving from state 2 to state 5, with and without a treatment offer

et al [13] in their study. But they are barely comparable since the data set of Stern et al is a whole lot bigger than my simulated sets.

5.3. Simulation of patients with start state 1

I simulate 1000 patients from the same Markov model in Figure 5.1 where all patients start in state 1, i.e. all patients start with no treatment and high CD4 count. Only 64 patients end up in state 5 which means they end up with either AIDS or death and 936 patients get censored.

Since all the patients start in the same state, there is no use in doing an analysis with the baseline covariates since they are all the same for every patient. I do two different analyses for each method:

(1) treatment as the only covariate
(2) treatment and time-updated covariates

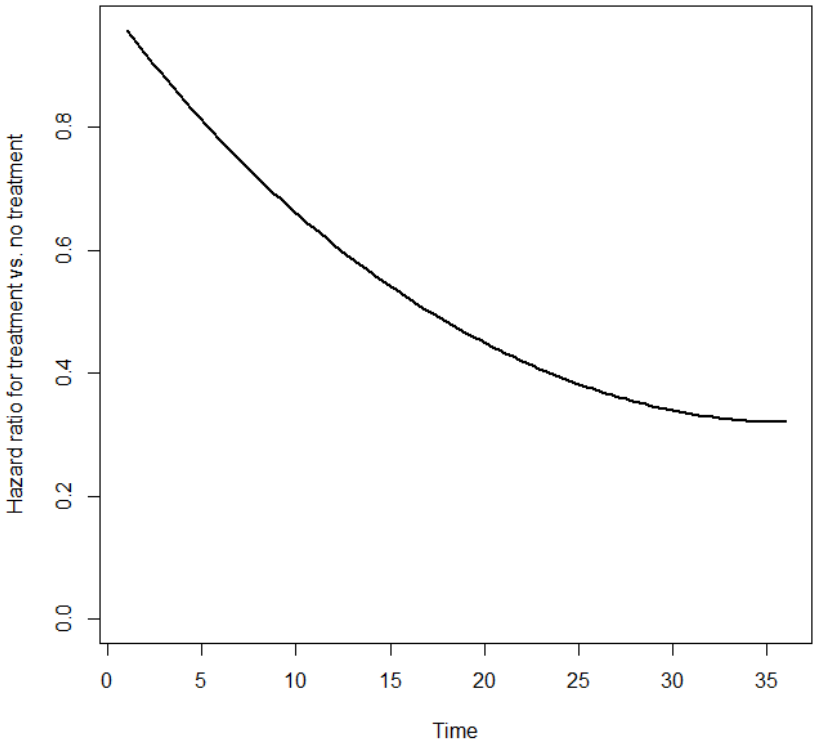

FIGURE 5.7. Hazard Ratio, start state 2

Analysis no.	**HR**	**SE**	**z**	**P-value**	95% **C.I**
(1)	0.9169	0.2026	−0.428	0.6680	[0.6164, 1.3640]
(2)	0.7431	0.2106	−1.410	0.1590	[0.4918, 1.1230]

TABLE 5.3. Unweighted Cox Analysis

5.3.1. Unweighted Cox Analysis. From the unweighted analyses in Table 5.3 the hazard of AIDS or death is reduced with 8% when the baseline covariates are used and reduced with 26% when the time-dependent covariates are used. The last estimate cannot be trusted because there is small variation in the CD4 count, and this gives a warning in R saying "Loglik converged before variable 2; beta may be infinite". There is too little variation in the cd4 data to estimate the treatment effect correctly. Maybe not so weird since everyone starts out in the same state.

5.3.2. Weighted Cox Analysis. It is not possible to make stable weights here since the baseline covariates are all the same.From the Logistic analysis in Table 5.4 we can see that the hazard for getting AIDS or dying is reduced with 69% for the patients on treatment. I don't

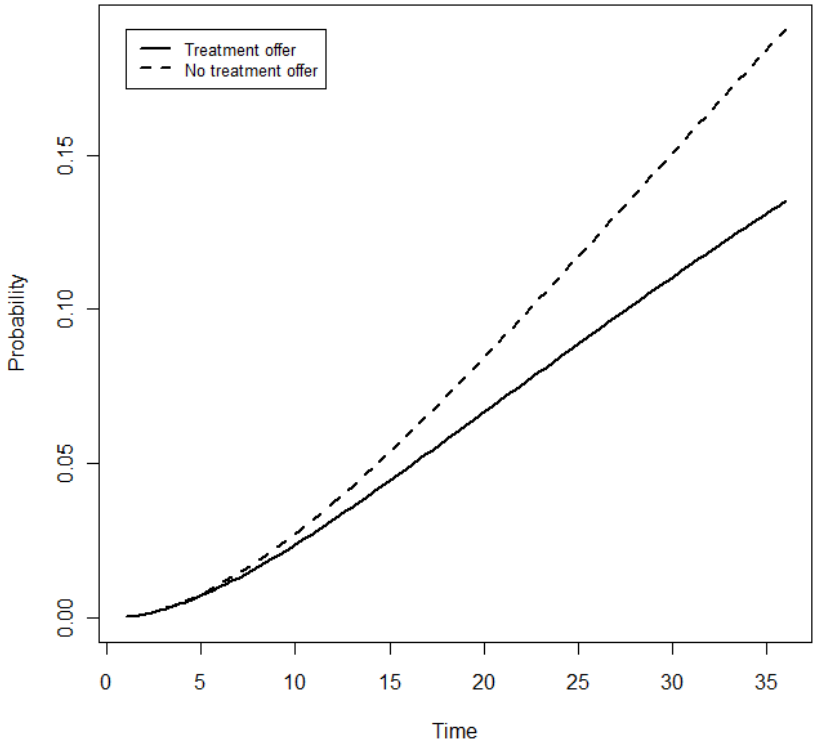

FIGURE 5.8. $p_{1,5}(t)$: Probabilities of moving from state 1 to state 5, with and without a treatment offer

Analysis	**HR**	**SE**	**z**	**P-value**	95% **C.I**
Weighted	0.6198	0.2060	−2.32179	0.0202	[0.4139, 0.9282]

TABLE 5.4. Results from the weighted analysis, start state 1

know how accurate these results are if I analyze the weights used for the analyses. The weights vary from 1 to 292. Some patients contribute to the set a huge amount of times compared to others. And the density of the weights is very skewed as can be seen from the histogram. But the patients all start out less sick and it then seems like the treatment is more effective.

5.4. Simulation of patients with start state 2

I simulated 1000 patients over three years where all the patients started in state 2. 550 patients ended up death or with AIDS and 450 did not experience the event before the study time was over, e.g. were censored.

5.4.1. Unweighted Cox Analysis.

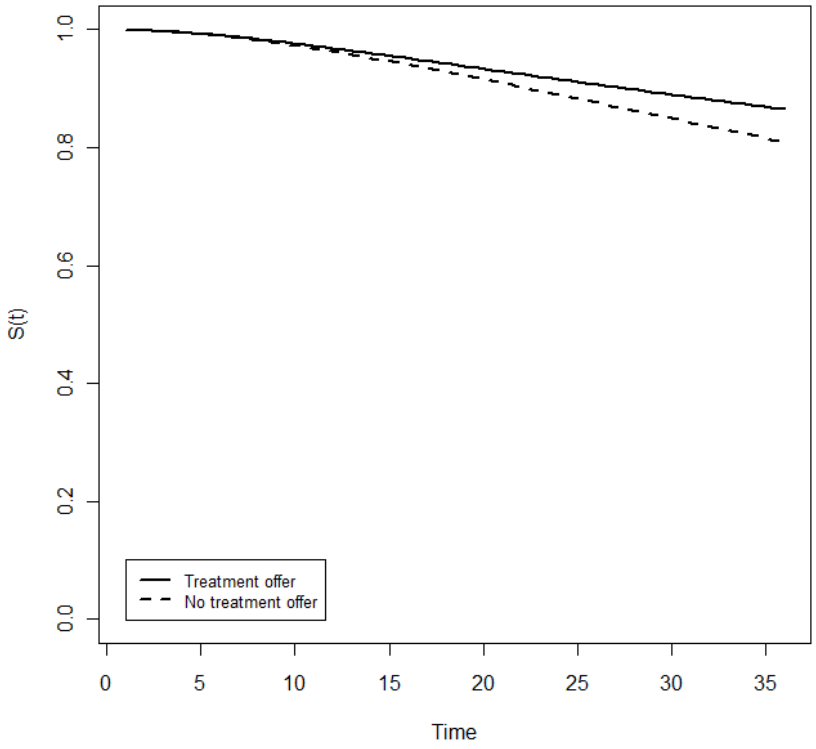

FIGURE 5.9. $S_{1,5}(t)$: Probabilities of not moving from state 1 to state 5, with and without a treatment offer

Analysis no.	**HR**	**SE**	**z**	**P-value**	95% **C.I**
(1)	0.5189	0.1001	-6.555	≤ 0.001	$[0.4265, 0.6314]$
(2)	0.6472	0.1084	-4.014	≤ 0.001	$[0.5234, 0.8004]$

TABLE 5.5. Unweighted Cox Analysis

Analysis	**OR/HR**	**SE**	**z**	**P-value**	95% **C.I**
Weighted	0.8555	0.1028	-1.518	0.1290	$[0.6993, 1.0465]*$

TABLE 5.6. Results from the weighted analysis, start state 2

5.4.2. Weighted Cox Analysis.

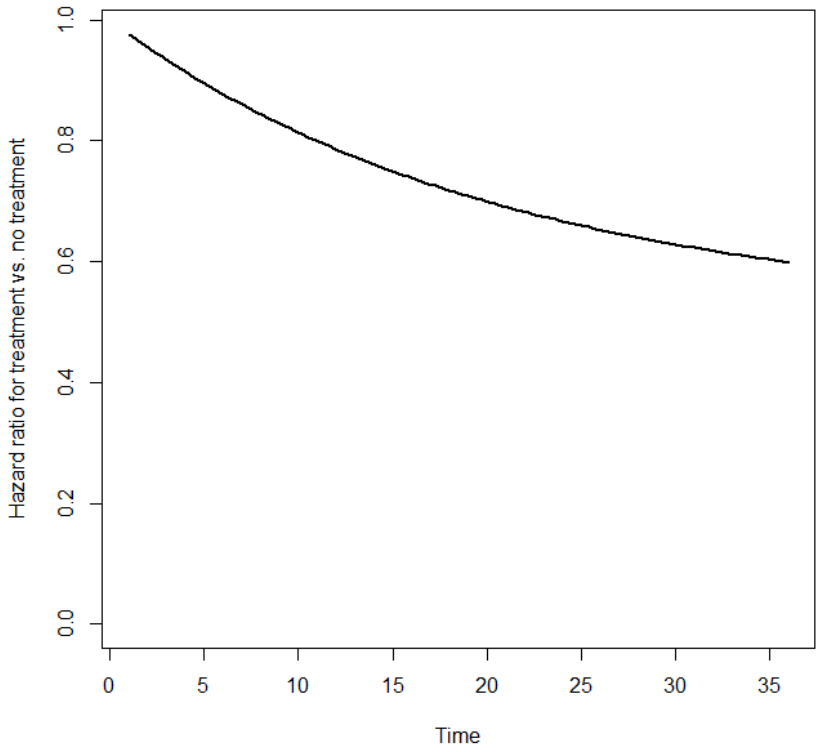

FIGURE 5.10. Hazard Ratio, start state 1

5.5. Simulation of patients with start states 1 and 2

I simulate 1000 new patients from the Markov model where one half start out in state 1 and the second half in state 2. They are followed up for a period of 3 years, and 295 patients end up in state 5 which means they end up with either AIDS or death. 705 get censored.

The three different analyses for a unweighted Cox and pooled logistic are done again.

Analysis no.	**HR**	**SE**	**z**	**P-value**	95% **C.I**
(1)	6.8841	0.1427	13.52	≤ 0.001	$[5.2040, 9.1060]$
(2)	3.5584	0.1561	8.131	≤ 0.001	$[2.6205, 4.8321]$
(3)	3.6866	0.1564	8.341	≤ 0.001	$[2.7132, 5.0093]$

TABLE 5.7. Unweighted Cox Analysis

5.5.1. Unweighted Cox Analysis. From the unweighted analyses in Table 5.7 the hazard of AIDS or death is from 3 to 6 times higher for the patients on treatment. These estimates are biased.

Analysis	**OR/HR**	**SE**	**P-value**	95% **C.I**	
Cox	2.9930	0.1350	8.1219	≤ 0.001	$[2.2973, 3.8994]$

TABLE 5.8. Results from the weighted analysis, start states 1 and 2

5.5.2. Weighted Cox Analysis.

5.6. Simulation of patients with start states 3 and/or 4

It makes no sense to do the analyses with patients who start in state 3 or 4, since they all start with treatment.

CHAPTER 6

Concluding Remarks

In the analyses for my thesis I show the use of the Marginal Structural Model. I estimate hazard ratios by using weighted Cox proportional hazards models, controlling for time-dependent confounding. The weights are calculated from the inverse of each patient's probability of the treatment history they actually had, given their covariate history. This gives me a weighted set where the treatment probability is unrelated to the time-dependent confounders. The confounders are controlled by the weights and not as covariates in the Cox models. With this I also avoid the problem that the confounders can be intermediate on the causal pathway from HAART to the outcome of AIDS or death.

The analyses done in my thesis have several restrictions. For example there has been no placebo-controlled randomized trial of HAART. The reason for this is of course because it wouldn't be ethically right to offer treatment to some patients and not to other patients when they all need treatment. And in many studies the follow-up time has been of a year or less. With this the effectiveness of HAART over several years is still unknown. The use of the MSM gives an estimate of the effect over several years, but there is maybe no way of knowing how good the estimation is because of the ethical restrictions. Another thing could be the use of CD4 count and viral load as markers for progression to AIDS or death. In my thesis I have assumed that these measures are good enough as markers...(kanskje det fins en artikkel som diskuterer dette...)

Standard methods for estimating the causal effect of treatment on AIDS or death will produce biased estimates:

(1) The crude estimate without control for confounding is biased because the subjects on HAART usually have a low CD4 count, and subjects with low count values have higher AIDS and death rates
(2) The estimate when controlling for baseline values such as the CD4 count at baseline will give biased results bacause the fact that the subjects starting on HAART had low CD4 counts is ignored
(3) Controlling for the time-dependent confounders such as the CD4 count will produce biased estimates because HAART will partly make the CD4 counts higher

References

[1] the free encyclopedia Wikipedia. Facts about HIV/AIDS. 2001. {http://www.thebody.com/content/art32981.html}(2011-09-14).

[2] AIDS Education & Training Centers National Resource Center. HIV classification: CDC and WHO staging systems. 2011. {http://aidsetc.org/aidsetc?page=cm-105_disease}(2011-09-12).

[3] AVERTing HIV and AIDS. AIDS and HIV information. 2010. {http://www.avert.org/hiv.htm}(2011-09-14).

[4] Staistics Norway. HIV-infeksjon etter smittemåte og AIDS, etter diagnoseår. 2010. {http://www.ssb.no/aarbok/tab/tab-129.html}(2011-09-14).

[5] J.M. Robins, M.Á. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[6] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[7] SHCS. Swiss HIV cohort study. 2000. {http://www.shcs.ch}(2011-09-14).

[8] L. A. Cupples, R. B. D'Agostino, K. Anderson, and W. B. Kannel. Comparison of baseline and repeated measure covariate techniques in the farmingham heart study. *Statistics In Medicine*, 7:205–218, 1988.

[9] R.B. D'Agostino, M.L. Lee, A.J. Belanger, L.A. Cupples, K. Anderson, and W.B. Kannel. Relation of pooled logistic regression to time dependent cox regression analysis: the framingham heart study. *Statistics in Medicine*, 9(12):1501–1515, 1990.

[10] O.O. Aalen, Ø. Borgan, and H.K. Gjessing. *Survival and event history analysis: a process point of view*. Springer Verlag, 2008.

[11] M.Á. Hernán, B. Brumback, and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5):561–570, 2000.

[12] Z. Fewell, M.A. Hernán, F. Wolfe, K. Tilling, H. Choi, and JA Sterne. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*, 4(4):402–420, 2004.

[13] J.A.C. Sterne, M.A. Hernán, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J.M. Robins, and M. Egger. Long-term effectiveness of potent antiretroviral therapy in

preventing aids and death: a prospective cohort study. *The Lancet*, 366(9483):378–384, 2005.

[14] StataCorp LP. Stata - data analysis and statistical software. 1996-2011. {http://www.stata.com}(2011-09-14).

[15] the free encyclopedia Wikipedia. Viral load. 2011. {http://www.en.wikipedia.org/wiki/Viral_load}(2011-12-06).

[16] the free encyclopedia Wikipedia. Hemoglobin. 2011. {http://www.en.wikipedia.org/wiki/Hemoglobin}(2011-12-06).

[17] O. O. Aalen, V. T. Farewell, D. De Angelis, N. E. Day, and O. N. Gill. A markov model for hiv disease progression including the effect of HIV diagnosis and treatment: Application to aids prediction in england and wales. *Statistics In Medicine*, 16:2191–2210, 1997.

[18] L. J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology*. Pearson, 2003.